

---

# On degeneracy control in overcomplete ICA

---

**Jesse A. Livezey \***

Department of Physics  
Redwood Center for Theoretical Neuroscience  
University of California, Berkeley  
Berkeley, CA 94720  
jesse.livezey@berkeley.edu

\*Indicates equal contribution.

**Alejandro F. Bujan \***

Redwood Center for Theoretical Neuroscience  
University of California, Berkeley  
Berkeley, CA 94720  
afbujan@berkeley.edu

**Friedrich T. Sommer**

Redwood Center for Theoretical Neuroscience  
University of California, Berkeley  
Berkeley, CA 94720  
fsommer@berkeley.edu

## Abstract

Understanding the effects of degeneracy control mechanisms when learning overcomplete representations is crucial for applying Independent Components Analysis (ICA) in machine learning and theoretical neuroscience. A number of approaches to degeneracy control have been proposed which can learn non-degenerate complete representations, however some of these methods can fall into bad local minima when extended to overcomplete ICA. Furthermore, they may have unintended side-effects on the distribution of learned basis elements, which may lead to a biased exploration of the data manifold. In this work, we identify and theoretically analyze the cause of these failures and propose a framework that can be used to evaluate arbitrary degeneracy control mechanisms. We evaluate different methods for degeneracy control in overcomplete ICA and suggest two novel approaches, one of which can learn highly orthonormal bases. Finally, we compare all methods on the task of estimating an overcomplete basis on natural images.

## 1 Introduction

Independent Components Analysis (ICA) is a technique for learning the underlying non-Gaussian and mutually independent sources  $S$  in a dataset  $X$  (Comon, 1994; Bell and Sejnowski, 1997). ICA is formulated as a linear generative model:

$$X = AS, \quad (1)$$

where  $A$  is usually referred to as the *mixing matrix*. In the complete case, the reconstruction of the original sources is possible as the mixing matrix can be inverted. The goal of ICA is then to find the *unmixing matrix*  $W$  such that the sources can be recovered,  $S = WX$  with  $W = A^{-1}$ . The unmixing matrix  $W$  can be obtained by solving the following constrained optimization problem:

$$\begin{aligned} \arg \min_W \quad & \sum_{i=1}^m \sum_{j=1}^k g(W_j x^{(i)}) \\ \text{s.t.} \quad & WW^T = I \end{aligned} \quad (2)$$

where  $g(\cdot)$  is termed the *contrast function* and is usually a softer version of the  $L_1$  norm like the  $\log(\cosh(\cdot))$ . The constraint  $WW^T = I$  prevents the bases from becoming degenerate (Hyvärinen and Oja, 1997).

This constrained optimization can be relaxed into an unconstrained one by adding a new cost  $C$  to the objective function (Le et al., 2011). This cost should then play the role of the constraint, preventing the co-alignment of the bases. The new unconstrained optimization problem becomes:

$$\arg \min_W C(W) + \lambda \sum_{i=1}^m \sum_{j=1}^k g(W_j x^{(i)}). \quad (3)$$

In addition to its use in machine learning, e.g. blind source separation, ICA learns filters which are similar to neuronal receptive fields found in early stages of visual cortex (V1), which gives insight into the connections between natural scene statistics and sensory learning in the brain.

Overcomplete versions of sparse coding and ICA (Lewicki and Olshausen, 1999; Hyvärinen, 2005; Le et al., 2011) have been proposed. Overcomplete representations have been used to improve classification (Hinton et al., 2006; Bengio et al., 2007; Coates et al., 2011) and learn more diverse features (Rehn and Sommer, 2007; Olshausen, 2013). ICA, as an alternative to sparse coding, has the advantage that the inference process is a linear transformation versus a *maximum a posteriori* (MAP) estimation or posterior estimation which often require computationally expensive methods. Overcomplete ICA lacks a natural mechanism for *explaining-away* between bases, which arises naturally during inference in sparse coding. *Degeneracy control* represents a way to incorporate an explaining-away-like effect into overcomplete ICA.

In overcomplete ICA, orthonormality can no longer be enforced for all bases, therefore some other form of degeneracy control (Hyvärinen et al., 1999) is needed to prevent bases from learning identical features, i.e. the angles between all pairs of bases should be pushed away from zero. Different methods for degeneracy control have been proposed: a quasi-orthogonality constraint (Hyvärinen et al., 1999), a reconstruction cost (referred to as  $L_2$  cost here) (Le et al., 2011), and a random prior cost (Hyvärinen and Inki, 2002) (see Methods for more details). However, a systematic analysis of the properties of these methods is still missing.

Here, we show that while the  $L_2$  cost has the correct behavior for a complete basis, in the overcomplete case, there are local minima which allow pairs of basis vectors to become arbitrarily close to each other. We introduce a theoretical framework for evaluating different degeneracy control costs, and propose two new costs, which we compare with previously proposed methods. Our first novel approach is the  $L_4$  cost on the difference between the identity matrix and the Gram matrix of the bases. The second method is a cost which we term the *Coulomb* cost because it is derived from the physics problem of the electrostatic repulsion between charged particles on the surface of a hyper-sphere.

In addition to preventing degeneracy, we show that these costs will influence the distribution of an overcomplete set of bases in the high dimensional space. We provide analytic and numerical methods for evaluating potential degeneracy controls and investigate their effect on the distribution of pairwise angles between bases. In this work, we apply this framework to the two novel methods as well as other methods from the literature. Finally, we evaluate the diversity of bases learned on natural images using different non-degeneracy costs.

## 2 Methods

### 2.1 Non-degeneracy constraints and costs

Here, we briefly describe previous methods for degeneracy control in overcomplete ICA.

#### 2.1.1 Quasi-orthogonality constraint

Hyvärinen et al. (1999) suggest a quasi-orthogonality update which approximates a symmetric Gram-Schmidt orthogonalization scheme for an overcomplete basis  $W$  and can be formulated as follows:

$$W \leftarrow \frac{3}{2}W - \frac{1}{2}WW^TW. \quad (4)$$

#### 2.1.2 Reconstruction cost and the $L_2$ cost

Le et al. (2011) propose adding a reconstruction cost to the ICA prior (RICA) as a form of degeneracy control, which they show is equivalent to a cost on the  $L_2$  norm of the difference between the Gram

matrix of the filters and an identity matrix for whitened data

$$C_{\text{RICA}} = \frac{1}{N} \sum_{ij} (X_j^{(i)} - \sum_{kl} W_{kj} W_{kl} X_l^{(i)})^2 \propto C_{L_2} = \sum_{ij} (\delta_{ij} - \sum_k W_{ik} W_{jk})^2 = \sum_{ij} (\delta_{ij} - \cos \theta_{ij})^2, \quad (5)$$

where  $W_{ij}$  is the component of the  $i$ th source for the  $j$ th mixture,  $X_j^i$  is the  $j$ th element of the  $i$ th sample,  $\theta$  is the angle between pairs of basis, and  $\delta_{ij}$  is the Kronecker delta.

### 2.1.3 Random prior cost

Hyvärinen and Inki (2002) use a prior on the distribution of pairwise angles to encourage quasi-orthonormality and non-degeneracy. The prior is derived from assuming that the distribution of pairwise angles for all basis vectors is the same as the distribution of pairwise angles for just two vectors<sup>2</sup>.

$$C_{\text{Random prior}} = -\log P(\cos \theta_{ij}) \propto -\sum_{i \neq j} \log(1 - \cos^2 \theta_{ij}) \quad (6)$$

## 2.2 Model implementation

All models were implemented in Theano (Theano Development Team, 2016) and trained using L-BFGS-B (Byrd et al., 1995) and the norm-ball projection (Le et al., 2011) to keep the bases normalized. A repository with code to reproduce the results will be posted online.

### 2.3 Fitting Gabor parameters

We fit the Gabor parameters (Ringach, 2002) to the learned bases using an iterative grid-search and optimization scheme which gave the best results on generated filters. The learned parameters were the center vector, planar-rotation angle  $\phi$ , phase, frequency, and envelope variances parallel and perpendicular to the oscillations. The outline of the procedure is listed in the supplement.

## 3 Theoretical results and novel costs

Non-degeneracy costs will influence the distribution of an overcomplete set of bases in the high dimensional space. The properties of a cost can be investigated theoretically by evaluating its behavior as a function of the pairwise angles between bases which are either nearly orthogonal or nearly parallel ( $|\cos \theta| \sim 0$  or  $1$  respectively). In order to isolate the effects of the degeneracy costs, we ignore the data-dependent ICA prior on the sources for the theoretical analysis.

### 3.1 Pathological degeneracy in the $L_2$ cost

For the  $L_2$  cost (Le et al., 2011), it can be shown that in the overcomplete case there exists a set of degenerate solutions in which the angle between pairs of bases becomes exactly zero. This is illustrated with a two dimensional, two times overcomplete example in figure 1. It can be shown that in this example, there are pathological minima, figure 1 (a), which can be continuously rotated into other “good” minima with no parallel bases, figure 1 (b). These solutions are equivalent in terms of the value of the cost, lie on a connected family of solutions, and yet one is degenerate.

In order to understand these minima, we evaluate the structure of the cost in the two dimensional example analytically. The global rotational symmetry of the cost allows us to parameterize all solutions with respect to one fixed basis element,  $\hat{x}$ , without loss of generality. The bases, shown in figure 1, are:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \cos \theta_1 \\ \sin \theta_1 \end{pmatrix}, \begin{pmatrix} \cos \theta_2 \\ \sin \theta_2 \end{pmatrix}, \begin{pmatrix} \cos \theta_2 + \theta_3 \\ \sin \theta_2 + \theta_3 \end{pmatrix}. \quad (7)$$

Setting  $\theta_1$  and  $\theta_3$  to  $\pi/2$ , i.e. creating two orthonormal bases, forms a ring of local minima as  $\theta_2$  is varied. This can be shown by computing the derivative of the cost and evaluating the eigenvalues of

<sup>2</sup>For both the random prior and the Coulomb cost (section 3.3), we regularize the costs and their derivatives near  $|\cos \theta| = 1$  by adding a small positive constant in the objective, i.e.  $1 - \cos \theta_{ij}^2 \rightarrow 1 + |\epsilon| - \cos \theta_{ij}^2$ .

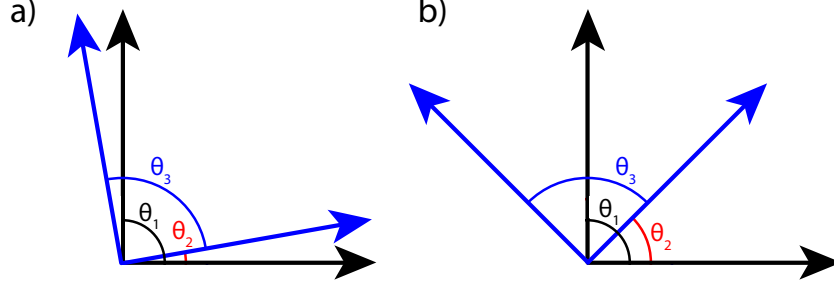


Figure 1: A pathological case in the  $L_2$  cost. The panels show two examples of overcomplete bases (black and blue arrows) which lie on a connected family of local minima for the  $L_2$  cost. **a)** shows a solution which has the same value of the cost as **(b)** for any  $\theta_2$  including the pathological solution  $\theta_2 \rightarrow 0$  and **b)** shows a potential good solution.

the Hessian of the cost at these points.

$$\begin{aligned}
 C_{L_2}(\theta_1, \theta_2, \theta_3)|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= 4 \\
 \frac{\partial C_{L_2}(\theta_1, \theta_2, \theta_3)}{\partial \vec{\theta}}|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= (0 \quad 0 \quad 0) \\
 \text{Eig.}(H_{L_2})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 0 \\ 8 \sin^2 \theta_2 \\ 8 \cos^2 \theta_2 \end{pmatrix}
 \end{aligned} \tag{8}$$

The first derivative is zero everywhere along this path, which shows that it is critical point. The first eigenvalue of the Hessian, 0, corresponds to moving in the  $\theta_2$  direction (see Supplement for details), which has no effect on the cost. The second two eigenvalues will be positive as long as  $\theta_2$  is not  $n\pi$  or  $(n + \frac{1}{2})\pi$  respectively. At the points where they are zero, the second derivatives vanish, but inspection of the cost along these axes show that they will be locally stable as the fourth derivative is positive.

As we show empirically in section 4.1, equivalent degenerate minima also exist in high dimensional overcomplete bases for the  $L_2$  cost. The simplest examples are subsets of  $N$  orthonormal bases in an multiple of  $N$  times overcomplete basis. Although these minima exist in high dimensions, it is unclear whether all local minima are of this type. We found that all local minima discovered from random initialization and numerical optimization have the same value of the cost as the degenerate solutions (data not shown).

Our numerical results, section 4.1, suggest that these pathological solutions are also present in the quasi-orthonormality constraint (Hyvärinen et al., 1999). In summary, we find that several proposed degeneracy control methods do not perform well in the overcomplete case which implies that there is a need for new forms of degeneracy control.

### 3.2 The $L_4$ cost as degeneracy control

We propose a novel degeneracy control cost termed the  $L_4$  cost, which transforms the degenerate minima into saddle points.

$$C_{L_4} = \sum_{ij} (\delta_{ij} - \sum_k W_{ik} W_{jk})^4 = \sum_{ij} (\delta_{ij} - \cos \theta_{ij})^4 \tag{9}$$

Following the same analysis as section 3.1, we show that the degenerate solutions are either reduced to saddle points at  $\theta_2 = n\frac{\pi}{2}$  or local minima at  $\theta_2 = (2n + 1)\frac{\pi}{4}$ , which correspond to good solutions.

$$\begin{aligned}
C_{L_4}(\theta_1, \theta_2, \theta_3)|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= 3 + \cos 4\theta_2 \\
\frac{\partial C_{L_4}(\theta_1, \theta_2, \theta_3)}{\partial \vec{\theta}}|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= (2 \sin 4\theta_2 \quad -4 \sin 4\theta_2 \quad -2 \sin 4\theta_2) \\
\text{Eig.}(H_{L_4})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 4(\cos 2\theta_2 - \cos 4\theta_2) \\ -2 \cos 2\theta_2 - 14 \cos 4\theta_2 - \dots \\ \dots \sqrt{2} \sqrt{34 - 2 \cos 2\theta_2 + \cos 4\theta_2 - 2 \cos 6\theta_2 + 33 \cos 8\theta_2} \\ -2 \cos 2\theta_2 - 14 \cos 4\theta_2 + \dots \\ \dots \sqrt{2} \sqrt{34 - 2 \cos 2\theta_2 + \cos 4\theta_2 - 2 \cos 6\theta_2 + 33 \cos 8\theta_2} \end{pmatrix} \quad (10)
\end{aligned}$$

### 3.3 The Coulomb cost as degeneracy control

We also propose a stronger cost, where the repulsion from large cosine distances is *Coulombic*: the Coulomb cost. Degeneracy control can then be related to the problem of characterizing the minimum energy states of  $M$  charged particles on a  $n$ -sphere, an open problem in electrostatics (Smale, 1998). The energy,  $E^{\text{Coulomb}}$ , of two charged point particles of the the same sign is proportional to the inverse of their distance,  $\vec{r}_{ij}$ :

$$E_{ij}^{\text{Coulomb}} \propto \frac{1}{|\vec{r}_{ij}|}. \quad (11)$$

To map this problem onto ICA, the cost should be made symmetric around  $\theta = \pi/2$  which can be accomplished by replacing  $\theta$  with  $2\theta$ , i.e.  $|r_{ij}| = \sqrt{1 - \cos^2 \frac{\theta_{ij}}{2}} \rightarrow \sqrt{1 - \cos^2 \theta_{ij}}$ . As a results, the Coulomb cost can be formulated as follows:

$$C_{\text{Coulomb}} = \sum_{ij} \frac{1}{\sqrt{1 - \cos^2 \theta_{ij}}} \quad (12)$$

We evaluate this cost numerically in section 4.1.

## 4 Experimental Results

### 4.1 Distribution of bases

In high dimensional spaces, it may be difficult to analytically understand the behavior of different degeneracy control mechanisms. In these cases, the behavior can be analyzed numerically by optimizing the different costs,  $C$ . To more systematically probe the effects of the different controls on the pairwise angle distributions, we use two different initializations of the bases. These are: a random uniform initialization, and a pathological initialization (as in section 3.1).

In the random uniform case, the bases are evenly initialized on the unit hyper-sphere. After the optimization, the quasi-orthogonality update (Hyvärinen et al., 1999) produces a distribution which is less uniform than random (figure 2 a). The other costs force the angles close to 90 degrees at the cost of producing distributions with longer tails. This trade-off is particularly pronounced for the  $L_2$  cost which peaks at 90 degrees but also has the longest tail towards small angles. The other 3 costs have shorter tails and have varying amounts of density near 90 degrees. Of all costs, the  $L_4$  cost distributes the angles most evenly which is reflected by its distribution having the narrowest width.

The pathological initialization shows how the quasi-orthogonality update and the  $L_2$  cost can fall into pathological minima. The pathological initialization tiles an orthonormal, complete basis two times and adds Gaussian noise to every basis element. Most bases start either close to either 90 or 0 degrees apart as shown in the two peaks in the initial distribution. The  $L_2$  cost is unable to move away from this solution unlike the other costs, which generate comparable solutions for all initializations.

In order to gain more insight into the qualitative differences in the distributions of angles, we analyze the behavior of the costs around  $\theta = 0$  and  $\theta = 90$  (figure 2c-d respectively). The gradient of the cost close to  $|\cos \theta| = 1$  is proportional to the force the angles feel to stay away from zero which will determine the tail of the angle distribution.

Taylor expanding all the costs near  $\cos \theta = 0$  reveals that all cost functions have non-zero second order terms except for the  $L_4$  cost which only has a fourth order term. Costs which have non-zero

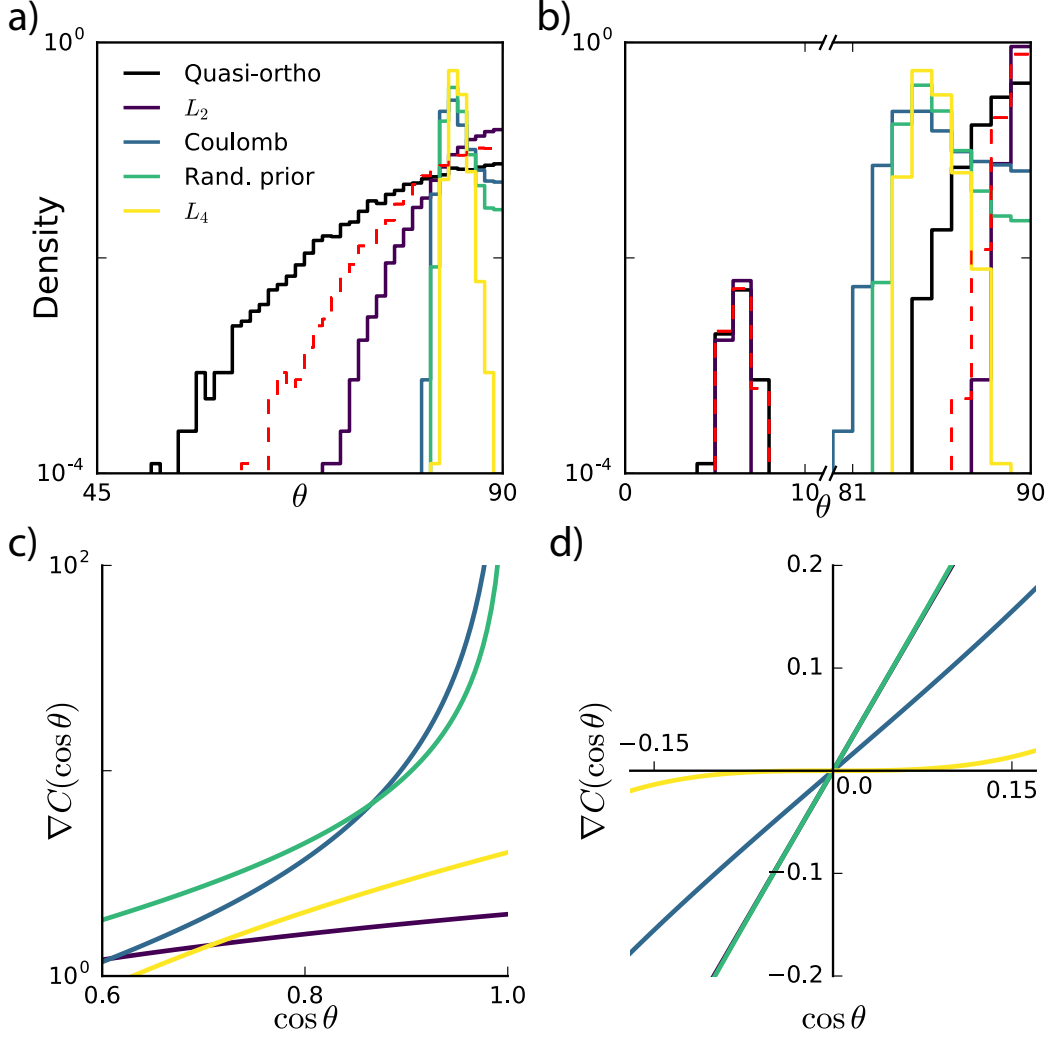


Figure 2: Analysis of different degeneracy control mechanisms. Color code is preserved across panels. **a)** Angle distributions obtained by numerically minimizing different cost functions  $C$  for a random initialization of the bases (see main text section 4.1). Red dotted line indicates the initial distribution of angles. **b)** Angle distributions obtained (as in (a)) with a pathological initialization of the bases. **c)** Gradient of the costs as a function of  $\cos \theta$  near  $\cos \theta \sim 1$ . **d)** Gradient of the costs as a function of  $\cos \theta$  near  $\cos \theta \sim 0$ .

second order terms will have gradients which scale linearly in  $\cos \theta$ , whereas the  $L_4$  cost's gradient grows slowly ( $\sim x^3$ ) near zero, as shown in figure 2d. These gradients affect the behavior of bases which are nearly orthogonal, namely gradients which scale linearly will encourage pairs of basis vectors to be more orthogonal at the expense of skewing the angle distribution towards small values. This leads to distributions of angles which are less uniform over all elements of the basis.

The effect of the gradient of the costs for intermediate values of  $\theta$  is difficult to analyze because in addition to forces coming from the costs, there are effective forces coming from the constraint that all of the basis vectors must live on a unit hyper-sphere. For instance, since the gradient of the random prior is larger than the Coulomb cost near zero, one would expect this to cause the distribution to be higher near 90 degrees. Surprisingly, this is not the case, which can be understood by taking into account that the gradients at smaller angles are also larger which can lead to a distribution with a higher peak and less mass near 90 degrees.

## 4.2 Experiments on natural images

We test whether the results obtained from evaluating the degeneracy control costs in isolation persist when the costs are used to train ICA models on natural images, i.e., equation 3. We train four times overcomplete ICA models on 8-by-8 whitened image patches at a fixed level of sparsity across costs. It is known that for natural images data sets, bases learned with ICA can be well-fit by Gabor filters (Bell and Sejnowski, 1997). Hence, we evaluate the distribution of the learned basis by inspecting the parameters obtained from fitting the bases to Gabor filters (see supplement for details).

Our results show that the distributions of angles from the trained ICA models are inline with the theoretical results from section 4.1 (figure 3 a). The  $L_2$  cost has a long tail compared to the other costs with the  $L_4$  having the largest minimum angle between bases. For the range of sparsities which were considered, the visual appearance of the bases is similar to results from previous ICA work and similar across costs ( $L_4$  bases are shown in figure 3 b). We also find that for the degrees of sparseness used in the experiments, the reconstruction errors are comparable across costs.

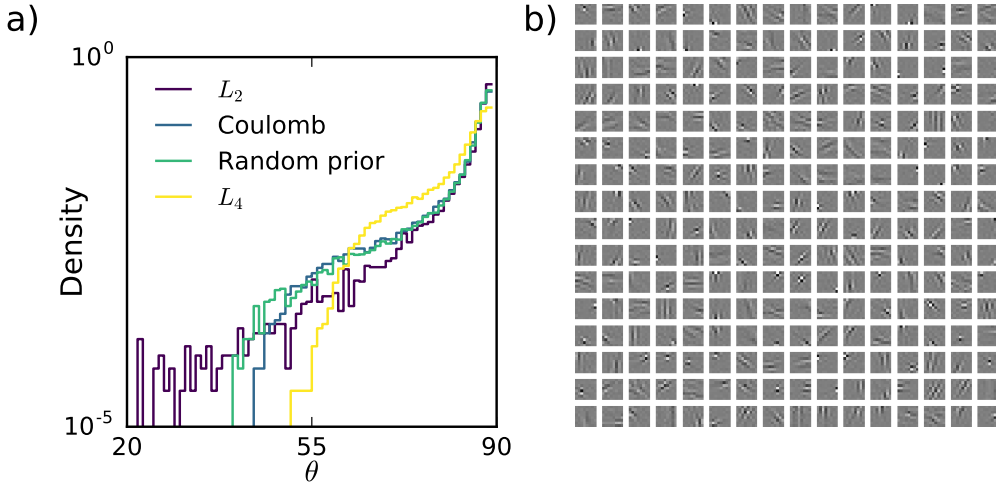


Figure 3: Results from fitting a four times overcomplete model on natural images. **a)** Angle distributions across costs for 8-by-8 natural image patches for a fixed value of sparsity (y-axis log scaled). **b)** Bases learned with the  $L_4$  cost.

We then visualize in more detail the tiling properties of the learned dictionaries. Figure 4 shows the summary of these results across costs. We find that for some parameters, the  $L_4$  cost has superior tiling properties. In particular, the distribution of angles and scales seems to better cover the range. The last column of (a) and (b) in figure 4 most clearly shows this difference. Altogether these results suggest that the use of a cost which encourages a more even distribution of basis vectors on the hyper-sphere has the potential to recover a larger diversity of sources from a natural images dataset.

## 5 Discussion

In overcomplete ICA, degeneracy control is needed to prevent bases from becoming co-aligned. As we have shown, the choice of degeneracy control costs will also have an influence on the distribution of bases elements. In this paper we suggest two novel costs which prevent degenerate solutions. We show the  $L_2$  cost and quasi-orthogonality update have pathological solutions which are not present when used in complete ICA.

In general, the choice of a model and cost function should take into account knowledge about the structure of the data on which it is being trained. For instance, natural images have translational and rotational invariances both locally and globally which may not be modeled well by a degeneracy cost which also promotes global heterogeneity. Subspace ICA methods (Hyvärinen and Hoyer, 2000) have been proposed to model these invariances and an overcomplete subspace-ICA model may benefit from a careful evaluation of different degeneracy control costs.

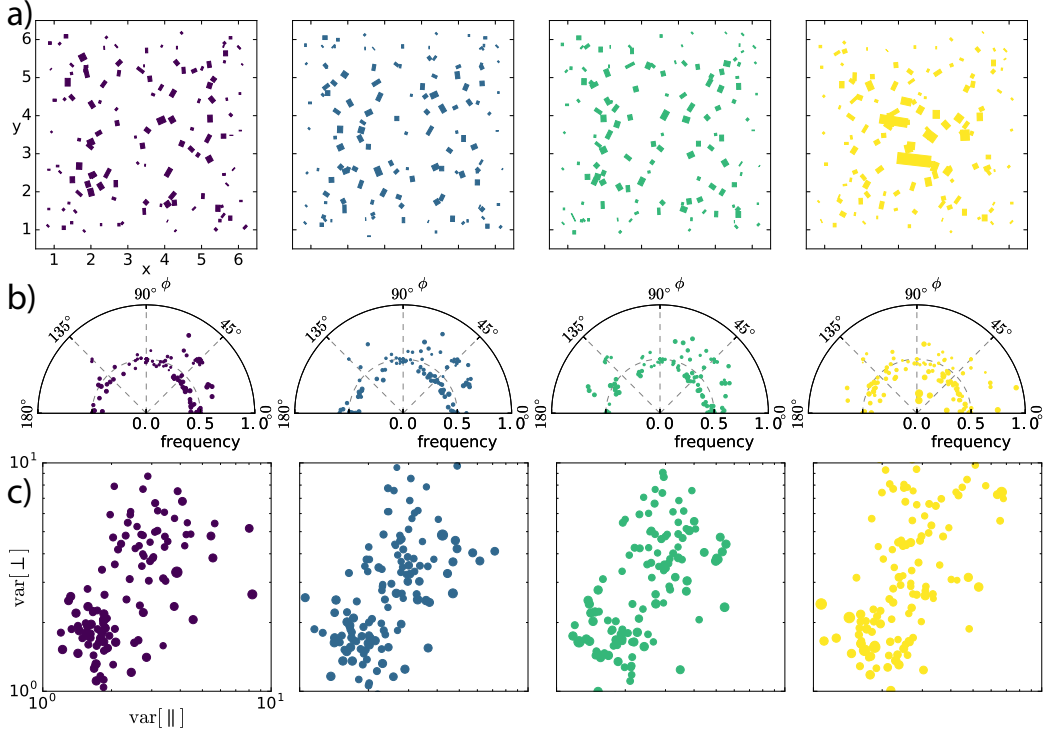


Figure 4: Results from fitting four times overcomplete ICA models on natural images with different costs. Color code as in previous figures. **a)** Rectangle position: center of Gabor fit in pixel coordinates, rectangle rotation: planar-rotation of Gabor, rectangle shape: envelope variances parallel and perpendicular to the oscillation axis. **b)** Polar plots of planar-rotation angle and spatial frequency. Marker size scales with envelope area. **c)** Log-scale plot of envelope variances parallel and perpendicular to the oscillation axis. Marker size scales with spatial frequency.

Ideally, the distribution of bases should only be influenced by the dataset they are being learned on. Practically, the random initialization of bases, degeneracy control costs, and optimization algorithm can bias the distribution of bases learned. Therefore, it is important to have a theoretical and computational framework to evaluate these biases.

## 6 Acknowledgements

We thank Yubei Chen, Alexander Anderson, and Kristofer Bouchard for their helpful discussions. JAL was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. JAL and AFB were supported by the Applied Mathematics Program within the Office of Science Advanced Scientific Computing Research of the U.S. Department of Energy under contract No. DE-AC02-05CH11231. FTS was supported by INTEL, the Kavli Foundation and the National Science Foundation (grants 0855272, 1219212, 1516527).



## References

- Bell, A. J. and Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision research*, 37(23):3327–3338.
- Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153.
- Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.
- Coates, A., Ng, A. Y., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics*, pages 215–223.
- Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.
- Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, pages 695–709.
- Hyvärinen, A., Cristescu, R., and Oja, E. (1999). A fast algorithm for estimating overcomplete ica bases for image windows. In *Neural Networks, 1999. IJCNN’99. International Joint Conference on*, volume 2, pages 894–899. IEEE.
- Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720.
- Hyvärinen, A. and Inki, M. (2002). Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 17(2):139–152.
- Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492.
- Le, Q. V., Karpenko, A., Ngiam, J., and Ng, A. Y. (2011). Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, pages 1017–1025.
- Lewicki, M. S. and Olshausen, B. A. (1999). Probabilistic framework for the adaptation and comparison of image codes. *JOSA A*, 16(7):1587–1601.
- Olshausen, B. A. (2013). Highly overcomplete sparse coding. In *IS&T/SPIE Electronic Imaging*, pages 86510S–86510S. International Society for Optics and Photonics.
- Rehn, M. and Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of computational neuroscience*, 22(2):135–146.
- Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463.
- Smale, S. (1998). Mathematical problems for the next century. *The Mathematical Intelligencer*, 20(2):7–15.
- Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.

## Appendix

### Minima analysis for the $L_2$ and $L_4$ costs for a 2-dimensional space

Here we show the full Hessian matrices, eigenvalues, and eigenvectors for the analysis in the main text sections 3.1 and 3.2.

#### $L_2$ cost

$$\begin{aligned}
C_{L_2}(\theta_1, \theta_2, \theta_3)|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= 4 \\
\frac{\partial C_{L_2}(\theta_1, \theta_2, \theta_3)}{\partial \vec{\theta}}|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= (0 \quad 0 \quad 0) \\
H(C_{L_2})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 4 & 0 & 4 \cos 2\theta_2 \\ 0 & 0 & 0 \\ 4 \cos 2\theta_2 & 0 & 4 \end{pmatrix} \\
\text{Eig.}(H_{L_2})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 0 \\ 8 \sin^2 \theta_2 \\ 8 \cos^2 \theta_2 \end{pmatrix} \\
\text{EVec.}(H_{L_2})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}
\end{aligned} \tag{13}$$

#### $L_4$ cost

$$\begin{aligned}
C_{L_4}(\theta_1, \theta_2, \theta_3)|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= 3 + \cos 4\theta_2 \\
\frac{\partial C_{L_4}(\theta_1, \theta_2, \theta_3)}{\partial \vec{\theta}}|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= (2 \sin 4\theta_2 \quad -4 \sin 4\theta_2 \quad -2 \sin 4\theta_2) \\
H(C_{L_4})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} -8 \cos 4\theta_2 & 8 \cos 4\theta_2 & 4(\cos 2\theta_2 + \cos 4\theta_2) \\ 8 \cos 4\theta_2 & -16 \cos 4\theta_2 & -8 \cos 4\theta_2 \\ 4(\cos 2\theta_2 + \cos 4\theta_2) & -8 \cos 4\theta_2 & -8 \cos 4\theta_2 \end{pmatrix} \\
\text{Eig.}(H_{L_4})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 4(\cos 2\theta_2 - \cos 4\theta_2) \\ -2 \cos 2\theta_2 - 14 \cos 4\theta_2 - \dots \\ \dots \sqrt{2} \sqrt{34 - 2 \cos 2\theta_2 + \cos 4\theta_2 - 2 \cos 6\theta_2 + 33 \cos 8\theta_2} \\ -2 \cos 2\theta_2 - 14 \cos 4\theta_2 + \dots \\ \dots \sqrt{2} \sqrt{34 - 2 \cos 2\theta_2 + \cos 4\theta_2 - 2 \cos 6\theta_2 + 33 \cos 8\theta_2} \end{pmatrix} \\
\text{EVec.}(H_{L_4})|_{\theta_1, \theta_3 = \frac{\pi}{2}} &= \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \\
&\begin{pmatrix} -1 \\ (\frac{\sqrt{2}}{8} \sqrt{\frac{2 \cos 2\theta_2 + \cos 4\theta_2 - \dots}{\dots 2 \cos 6\theta_2 + 33 \cos 8\theta_2 + 34 - \dots}} \dots \\ \dots - 2 \cos 2\theta_2) \sec 4\theta_2 + \frac{1}{4} \\ 1 \end{pmatrix}, \\
&\begin{pmatrix} -1 \\ \frac{1}{4} - (2 \cos \frac{1}{4} \theta_2 + \dots \\ \dots \frac{\sqrt{2}}{8} \sqrt{\frac{-2 \cos 2\theta_2 + \cos 4\theta_2 - 2 \cos 6\theta_2 + \dots}{\dots 33 \cos 8\theta_2 + 34}} \dots) \sec 4\theta_2 \\ 1 \end{pmatrix}
\end{aligned} \tag{14}$$

### Minima analysis for high dimensions

For an  $N$  dimensional space with an  $M$  (integer) times overcomplete basis, the pathological case is  $M$  orthonormal subsets, each its own orthonormal basis. We label our bases into the sequential

subsets of orthonormal subsets by label  $W_1, \dots, W_N, \dots, W_{2N}, \dots, W_{M \times N}$ . So, bases  $W_1$  through  $W_N$  form a full-rank orthonormal basis and this basis is tiled  $M$  times.

We show that starting from the pathological initialization, applying a rotation to one set of bases leaves the value of the cost unchanged, and we also show that this initialization is a critical point of the cost.

### Symmetry under rotations

We can show that the  $L_2$  cost is invariant to rotations applied to any basis subset. This shows that configurations where no basis are exactly aligned and configurations where a large number of bases are exactly aligned can have the same value of the cost. However, it does not imply that these configurations are local minima of the cost.

Consider the following partition of the bases: partition  $\mathcal{A}$  is the first orthonormal set, i.e. bases  $W_1$  through  $W_N$ , and partition  $\mathcal{B}$  the remainder of the bases, i.e. bases  $W_{N+1}$  through  $W_{M \times N}$ . If a rotation is applied to  $\mathcal{A}$ , only terms in the cost between elements of  $\mathcal{A}$  and  $\mathcal{B}$  will change. It is straightforward to show that the terms in the cost that have both elements within  $\mathcal{A}$  or both within  $\mathcal{B}$  are constant since the rotation does not alter the relative pairwise angles.

For some element  $W_i \in \mathcal{B}$ , we can write down the terms in the cost which contain itself and elements from  $\mathcal{A}$ :

$$C_{\mathcal{A}, W_i} = \sum_{W_j \in \mathcal{A}} (W_j^T W_i)^2 + (W_i^T W_j)^2 = 2 \sum_{W_j \in \mathcal{A}} \left( \text{Proj}_{W_j}(W_i) \right)^2 = 2|W_i|^2. \quad (15)$$

Since the  $W_j \in \mathcal{A}$  remain an orthonormal basis under a rotation, the sum of the projections-squared is just the  $L_2$  norm-squared of  $W_i$  which is constant. Since this is true for every  $W_i \in \mathcal{B}$ , the entire cost is constant under this rotation.

### Critical point analysis

We can also show that this initialization is a critical point of the cost for all dimensions and integer overcompleteness. This initialization is symmetric under the permutation of basis element labels, so showing that this initialization is a critical point for perturbations to one basis element shows that it is a critical point for perturbations to all elements.

Consider an infinitesimal rotation operator in  $S^2$ ,  $R = I + \epsilon G$ , expanded to first order for small  $\epsilon$  where  $G$  is the generator of the rotation  $R$  and  $I$  is the identity. Showing that the  $L_2$  cost has no first order dependence on  $\epsilon$  when the rotation is applied to a specific element,  $W_i$  shows that this initialization is a critical point. Since only  $W_j$  is being perturbed, we only need to consider terms in the cost which depend on  $W_i$ . We partition the terms into one set  $\parallel$ , which contains all bases which are parallel to  $W_i$  and a second set  $\perp$ , which contains all bases which are perpendicular to  $W_i$ :

$$\begin{aligned} C_{W_i}(\epsilon) &= \sum_{W_j \in \parallel} (W_j^T (I + \epsilon G) W_i)^2 + (((I + \epsilon G) W_i)^T W_j)^2 \\ &\quad + \sum_{W_j \in \perp} (W_j^T (I + \epsilon G) W_i)^2 + (((I + \epsilon G) W_i)^T W_j)^2. \end{aligned} \quad (16)$$

This can be simplified using the identity  $G^T = -G$  as  $G$  is a skew-symmetric matrix. Grouping the terms by their  $\mathcal{O}(\epsilon)$  dependence gives the following expression:

$$\begin{aligned} C_{W_i}(\epsilon) &= \sum_{W_j \in \parallel} 2\epsilon (W_j^T I W_i)(W_j^T G W_i) - 2\epsilon (W_i^T I W_j)(W_i^T G W_j) \\ &\quad + \sum_{W_j \in \perp} 2\epsilon (W_j^T I W_i)(W_j^T G W_i) - 2\epsilon (W_i^T I W_j)(W_i^T G W_j) \\ &\quad + \text{const.} + \mathcal{O}(\epsilon^2 + \dots) \\ &= 0 + \text{const.} + \mathcal{O}(\epsilon^2 + \dots). \end{aligned} \quad (17)$$

It is useful to note that this equality does not rely on the cancellation of the two pairs of terms, but, in fact, all four terms are exactly zero. The first two terms are exactly zero because  $\langle x, Gx \rangle = 0$  for a

skew-symmetric matrix,  $G$  (can be seen by applying skew-symmetric identity above.) The second two terms are zero because  $W_i$  and  $W_j$  are perpendicular and the first terms in the products are zero.

### **Fitting Gabor kernels**

The procedure for finding the best Gabor kernel parameters was to save the parameter set with best mean-squared error with the following trials:

1. for different initial widths, fit the center vector for the envelope to the absolute value of the basis after blurring it,
2. for different rotations and frequencies, numerically optimize the rotation, phase, and frequency of the Gabor
3. for the best fits from above, re-optimize the centers, widths, and phases.

A repository with code to reproduce the results will be posted online.